NSAM -929

THE K-COEFFICIENT,

A PEARSON-TYPE SUBSTITUTE FOR THE CONTINGENCY COEFFICIENT

Robert J. Wherry, Jr. and Norman E. Lane

May 1965

Research Report

THE K-COEFFICIENT,

A PEARSON-TYPE SUBSTITUTE FOR THE CONTINGENCY COEFFICIENT

Robert J. Wherry, Jr. and Norman E. Lane

Bureau of Medicine and Surgery
Project MR005.13-3003
Subtask 1 Report No. 43

Approved by

Captain Ashton Graybiel, MC, USN
Director of Research

Released by

Captain H. C. Hunley, MC, USN
Commanding Officer

25 May 1965

U. S. NAVAL SCHOOL OF AVIATION MEDICINE
U. S. NAVAL AVIATION MEDICAL CENTER
PENSACOLA, FLORIDA

SUMMARY PAGE

THE PROBLEM

Determining the relationship between two categorical or qualitative variables has previously been possible only through use of C, the contingency coefficient. C has several distinct disadvantages, among them its lack of comparability to the Pearson product-moment correlation measures of relationship. This lack of comparability frequently creates problems in the interpretation of contingency coefficients and the generalization of results.

FINDINGS

A technique is described which provides an extension of the Pearson correlation equation to the case of two categorical variables. Through canonical weighting, maximal scale values are assigned to categories of the qualitative variables. These scale values create an optimally weighted composite for each categorical variable, and the correlation between composites is the "K" coefficient, which avoids many of the disadvantages of C, and allows the computation of a product-moment relationship between two categorical variables.

INTRODUCTION

Many problems in psychology must of necessity deal with categorical data. In the past, when an investigator has had two categorical variables, and wished to express the relationship between them, his usual recourse was to compute C, the contingency coefficient. It is well documented that C has certain disadvantages. For example, the magnitude of C cannot be interpreted as indicating the same degree of relationship as an ordinary Pearson correlation coefficient. This is due, in part, to built-in upper limits of C. When the number of columns (h) equals the number of rows (g), then the upper limit of $C = \sqrt{(g-1)/g}$. The problem of interpreting C is complicated even further when $g \neq h$; for such cases the upper limit of C is unknown. Even for a 2 x 2 contingency table the magnitude of C does not equal the magnitude of φ, which is a Pearson equation and which can readily be computed.

PROCEDURE

THE CORRELATION RATIO

In attempting to find a Pearson-type substitute for the contingency coefficient, certain guide lines seemed evident. First, as mentioned above, the technique should, for a 2 x 2 table, degenerate into the four-fold contingency coefficient (φ). Secondly, the technique for a 2 x g table should yield the same result as if we had assigned a value of "1" to row 1, a "0" to row 2, and solved the problem by means of the correlation ratio equation.

The raw score form of the Pearson product-moment equation is

$$r_{XY} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - \Sigma Y^2}} \qquad (1)$$

Wherry, Sr. (1) pointed out in 1944 that if one had categorical data for variable X and interval data for variable Y, and if one assigned the mean Y-score of those individuals in category i as their X-score (these are known as the Wherry weights), i.e., $X_i = \bar{Y}_i = \Sigma^{n_i} Y/n_i$, then the Pearson equation could be shown to be equivalent to the correlation ratio equation. Using the Wherry weights, certain values in Eq. (1) can be rewritten. Thus,

$$\Sigma XY = \Sigma^{g}\Sigma^{n_i}(Y\bar{Y}_i) = \Sigma^{g}\Sigma^{n_i}(\bar{Y} \cdot \Sigma^{n_i} Y/n_i) = \Sigma^{g}\left[\frac{(\Sigma^{n_i} Y)^2}{n_i}\right] \qquad (2)$$

$$\Sigma X = \Sigma^{g}\Sigma^{n_i}\bar{Y}_i = \Sigma^{g} n_i \cdot \bar{Y}_i = \Sigma^{g} n_i \Sigma^{n_i} Y/n_i = \Sigma^{g}\Sigma^{n_i} Y = \Sigma^{N} Y \quad , \text{ and} \qquad (3)$$

$$\sum_{N} X^2 = \sum_{g} \sum_{n_i} Y_i^2 = \sum_{g} \left[n_i \cdot \left(\frac{\sum_{n_i} Y}{n_i} \right)^2 \right] = \sum_{g} \left[(\sum_{n_i} Y)^2 / n_i \right] \quad (4)$$

[Note that Eqs. (2) and (4) are equal.]

Substituting Eqs. (2), (3), and (4) into Eq. (1), we obtain

$$r_{XY} = \frac{N \sum_{g} (\sum_{n_i} Y)^2 / n_i - (\sum_{N} Y)^2}{\sqrt{N \sum_{g} (\sum_{n_i} Y)^2 / n_i - (\sum_{N} Y)^2} \sqrt{N \sum_{N} Y^2 - (\sum_{N} Y)^2}} \quad (5)$$

But the numerator is the square of the left term in the denominator; therefore,

$$r_{XY} = \frac{\sqrt{N \sum_{g} (\sum_{n_i} Y)^2 / n_i - (\sum_{N} Y)^2}}{\sqrt{N \sum_{N} Y^2 - (\sum_{N} Y)^2}} = \sqrt{\frac{N \sum_{g} (\sum_{n_i} Y)^2 / n_i - (\sum_{N} Y)^2}{N \sum_{N} Y^2 - (\sum_{N} Y)^2}} \quad (6)$$

Equation (6) is, of course, the correlation ratio equation for use when Y is a continuous variable and X is a categorical variable with g categories. The fact that the correlation ratio is derivable from the Pearson equation is unfortunately neglected in most statistical texts.

THE "K" COEFFICIENT

Wherry, Sr., in the article mentioned above, was primarily interested in development of a technique for using qualitative data in multiple regression techniques. An alternative method for using qualitative data in multiple regression studies has recently gained prominence in the literature. This technique is to create a dichotomous variable for each category of X. Such variables have been referred to as "categorical predictor variables" by Bottenberg and Ward (2) and as "pseudo dichotomous variables" by Wherry, Jr. (3). Thus, if an individual is in the first category of X, he receives a "1" on the first created dichotomous variable and a "0" on all other dichotomous variables representing variable X. If he is in the second category of X, he receives a "1" on the second created dichotomous variable and a "0" on all the other created dichotomous variables, et cetera.

If g dichotomous variables are created as stated above and these variables are used to predict variable Y, a continuous variable, the multiple correlation ($R_{Y \cdot X_1 X_2 \ldots}$) will also equal η and the raw score beta weights will equal the Wherry weights mentioned above. It is also of interest to point out that the shrunken multiple correlation R will equal Kelly's epsilon (ϵ), the unbiased estimate of η.

Thus, we may recognize that there exists a method of assigning numbers not covered in the usual discussions of interval, ordinal, and nominal scaling techniques. This new technique we will refer to as "maximal" scaling, because it is the assignment of the set of numbers to a set of mutually exclusive categories which will maximize the Pearson equation.

If one had two categorical variables, X (with g categories) and Y (with h categories), one could, as stated above, create a set of g dichotomous variables to represent X and a set of h dichotomous variables to represent Y. If one could then simultaneously solve for the optimal ("Maximal") weights for the h variables and the optimal weights for the g variables, so as to maximize the Pearson equation, one could have a statement of the magnitude of relationship between two categorical variables which is a Pearson-type of correlation coefficient. Fortunately, once one has computed the interrelationships among the two sets of dichotomous variables, Hotelling's (4) technique of canonical correlation can be used to solve for the maximal weights. The resulting Pearson coefficient will be referred to as the "K-coefficient." The use of the canonical correlation technique for scoring categorical data was first mentioned by Fisher (5) in 1938, and has also been reported elsewhere (6). The discussions of the technique are usually couched in terms of matrix algebra. Because both matrix algebra and canonical analysis are not well known, relatively few persons would be able to compute a K-coefficient even if they had the above references available. In fact, even if one understands the canonical technique, several modifications are necessary for handling two sets of categorical variables. For the above reasons it seems desirable to set forth the necessary steps in obtaining a K-coefficient.

AN EXAMPLE

Let us assume we desire to find the K-coefficient for the 7 x 8 contingency table shown in Table I. The contingency table should be arranged so that the numbers of rows (g) is equal to or less than the number of columns (h).

Step 1.

Compute n, N-n, and $\sqrt{n\,(N-n)}$ for each row and column of the contingency table as shown in Table I.

Step 2.

Compute a g by g+h intercorrelation matrix using the equation

$$r_{ij} = \frac{N \cdot c_{ij} - n_i \cdot n_j}{\sqrt{n_i\,(N-n_i)}\ \sqrt{n_j\,(N-n_j)}} \qquad (9)$$

where c_{ij} = the cell entry in Table I if i is an X-variable and j is a Y-variable, otherwise let $c_{ij} = 0$.

Accuracy should be maintained to the sixth and seventh decimal in computing the matrix. It is not necessary to compute the intercorrelations among the Y-variables. The results of step 2 are shown in Table II.

Table 1

A 7 x 8 Contingency Table

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	$n_{i\cdot}$	$N-n_{i\cdot}$	$\sqrt{n_{i\cdot}(N-n_{i\cdot})}$
X_1	3	1	22	0	5	7	0	6	44	148	80.697
X_2	0	2	0	1	0	0	0	0	3	189	23.812
X_3	0	7	13	3	1	7	4	19	54	138	86.325
X_4	3	0	14	0	3	1	0	1	22	170	61.156
X_5	1	2	13	0	2	5	2	9	34	158	73.294
X_6	0	1	0	1	0	1	1	1	5	187	30.578
X_7	1	2	10	0	1	5	1	10	30	162	69.714
$n_{\cdot j}$	8	15	72	5	12	26	8	46	N = 192 ✓	1152 = N(h-1) ✓	
$N-n_{\cdot j}$	184	177	120	187	180	166	184	146		1344 = N(g-1) ✓	
$\sqrt{n_{\cdot j}(N-n_{\cdot j})}$	38.367	51.527	92.952	30.578	46.476	65.696	38.367	81.951			

Table II

The Intercorrelations Among the g X-Variables and The Intercorrelations of the g X-Variables and the h Y-Variables

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
X_1	1.00000	-.06870	-.34108	-.19615	-.25293	-.08916	-.23463	-.07235	-.11255	.14078	-.08916	.11519	.03773	-.11369	-.13186
X_2		1.00000	-.07881	-.04532	-.05844	-.02060	-.05422	-.02627	.27630	-.09759	.24309	-.03253	-.04986	-.02627	-.07072
X_3			1.00000	-.22503	-.29018	-.10228	-.26919	-.13043	.12005	-.17348	.11592	-.11366	-.01058	.10145	.16454
X_4				1.00000	-.16668	-.05882	-.15481	.17048	-.10472	.19421	-.05882	.10977	-.09458	-.07501	-.16361
X_5					1.00000	-.07585	-.19962	-.02645	-.03336	.00705	-.07585	-.00705	.01578	.03982	.02730
X_6						1.00000	-.07037	-.03410	.07426	-.12666	.17861	-.04222	.03086	.12956	-.01516
X_7							1.00000	-.01795	-.01837	-.03704	-.07037	.05185	-.03930	-.01795	.09452

Table III

The h-1 Diagonal Factors Which Represent the X-Space, and The Relationships of the g X-Variables and the h Y-Variables to the Diagonal Factors, ($a_{f_i Y_i}$'s)

Factor	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
f_1	1.00000	-.06870	-.34108	-.19615	-.25293	-.08916	-.23464	-.07235	-.11255	.14078	-.08916	.11519	.03773	-.11369	-.13186
f_2	.00000	.99764	-.10248	-.05894	-.07600	-.02679	-.07050	.02135	.26920	-.08813	.23753	-.02468	-.04738	-.03416	-.07997
f_3	.00000	.00000	.93443	-.31888	-.41120	-.14495	-.38146	-.11552	-.11692	-.14393	.11757	-.08230	-.00275	.06332	.11918
f_4	.00000	.00000	.00000	.92540	-.38048	-.13412	-.35296	.15839	-.07960	.18450	-.02682	.11311	-.09817	-.08551	-.16878
f_5	.00000	.00000	.00000	.00000	.78511	-.26884	-.70751	.00126	-.03003	-.05983	-.05377	.03746	-.02135	.00252	-.03481
f_6	.00000	.00000		.00000	.00000	.93812	-.41148	-.02492	.07423	-.10287	.18762	-.02057	.01455	.12461	-.04667

Step 3.

Perform a diagonal factor analysis (7) extracting not more than g-1 factors. It turns out that a maximum of g-1 factors are needed to explain all the variance of the g X-variables since each of the X-variables is mutually exclusive and mutually exhaustive. That is, if a person is found in one row, he cannot be found in any other row, and if a person is included in the analysis, he must be found in one of the rows. If, by chance, the proportions of cases found in each column are identical for two rows, the rows should be combined since, from a predictive standpoint, there is no difference between the two categories.

The results of the diagonal factor analysis are shown in Table III.

Step 4.

Using the h loadings of the Y-variables on the g-1 diagonal factors, obtain the interrelationships of the Y-variables in the X-space. To obtain this matrix, use the equation

$$r'_{ij} = \sum_{s=1}^{f_{g-1}} (a_{f_s Y_i} \cdot a_{f_s Y_j}) \qquad (10)$$

where

r'_{ij} is the relationship of the i^{th} Y-variable to the j^{th} Y-variable in the X-space, and

$a_{f_s Y_i}$ is the diagonal factor loading of the i^{th} Y-variable on the s^{th} diagonal factor.

The results of step 4 are shown in Table IV.

Table IV

The Interrelationships of The Y-Variables in the X-Space, (r'_{ij}'s)

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
Y_1	.04475	-.04189	.06056	-.03410	-.03684	-.01188	-.03146	-.04721
Y_2		.11155	-.08051	.10540	-.04088	-.00779	.02698	.01826
Y_3			.09650	-.07787	.05546	-.01101	.05055	-.05709
Y_4				.11701	-.03471	-.00843	.03500	.00442
Y_5					.03527	-.00646	.02961	-.04246
Y_6						.01398	-.00731	.01512
Y_7							.04095	.03380
Y_8								.06986

Step 5.

Perform a Principal Axis factor analysis (8) extracting only the first Principal Axis. This step obtains the vector through the X-space which will explain a maximum amount of the Y variance explainable with only one vector. The Principal Axis loadings (a_{PY_i}'s) are shown in Table V.

Table V

The Principal Axis Loadings of the Y-Variables, (a_{PY_i}'s)

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
P.A.	-.18127	.29501	-.30391	.28791	-.16924	.01475	.15021	.15049

Step 6.

Obtain the relationships ($a_{f_m P}$'s) of the Principal Axis to each of the h-1 diagonal factors by solving g-1 set of equations of the type

$$a_{f_I Y_1} \cdot a_{f_I P} + a_{f_{II} Y_1} \cdot a_{f_{II} P} + \dots + a_{f_{g-1} Y_1} \cdot a_{f_{g-1} P} = a_{P Y_1},$$

$$a_{f_I Y_2} \cdot a_{f_I P} + a_{f_{II} Y_2} \cdot a_{f_{II} P} + \dots + a_{f_{g-1} Y_2} \cdot a_{f_{g-1} P} = a_{P Y_2}$$

$$\dots \quad + \dots \quad + \dots + \quad \dots \quad = \dots$$

$$a_{f_I Y_{g-1}} \cdot a_{f_I P} + a_{f_{II} Y_{g-1}} \cdot a_{f_{II} P} + \dots + a_{f_{g-1} Y_{g-1}} \cdot a_{f_{g-1} P} = a_{P Y_{g-1}}. \quad (11)$$

Notice that only g-1 of the available h Y-variables are used in the above simultaneous solution. A consistent solution will be obtained with any g-1 of the Y-variables used. For simplicity the first g-1 set was chosen. Table VI shows the initial values for the simultaneous equations as obtained from Tables III and V, and the relationship of the Principal Axis to the diagonal factors. As a computational check, the sum of the squares of the $a_{f_m P}$'s should equal 1.0 (within rounding error). This demonstrates that the Principal Axis Vector (obtained in Step 5) is wholly within the X-space, and must therefore be completely predictable from the g-1 X-Variables.

Table VI

The Beginning Values for the g-1 Simultaneous Equations to be Solved to Obtain the Relationship of the Principal Axis to the Diagonal Factors. Final Solution is Shown in the Row Labeled $a_{f.p}$

Coefficients for							
$a_{f_I p}$	$a_{f_{II} p}$	$a_{f_{III} p}$	$a_{f_{IV} p}$	$a_{f_V p}$	$a_{f_{VI} p}$		a_{pY_1}
-.07235	-.02135	-.11552	.15839	.00126	.02492	=	-.18127
-.11255	.26920	.11692	-.07960	-.03003	.07423	=	.29501
.14078	-.08813	-.14393	.18450	.05983	-.10287	=	-.30391
-.08916	.23753	.11757	-.02682	-.05377	.18762	=	.28791
.11519	-.02468	-.08230	.11311	.03746	-.02057	=	-.16924
.03773	-.04738	-.00275	-.09817	-.02135	.01455	=	.01475
$a_{f.p}$ -.46213	.44628	.47222	-.47346	-.14697	.34421		

$$\sum_{m=f_I}^{f_{g-1}} a^2_{f_s p} = .99997$$

Step 7.

Using the loadings of the Principal Axis on the diagonal factors as the criterion variable, solve the g-1 simultaneous equations necessary to obtain the standard-score beta weights of the g-1 X-variables used as the basis for the diagonal factors. The g-1 set of simultaneous equations for step 7 is of the form

$$\begin{aligned}
a_{f_I x_1} \cdot \beta_{x_1} + a_{f_I x_2} \cdot \beta_{x_2} + \ldots + a_{f_I x_{g-1}} \cdot \beta_{x_{g-1}} &= a_{f_I p}, \\
a_{f_{II} x_2} \cdot \beta_{x_2} + \ldots + a_{f_{II} x_{g-1}} \cdot \beta_{x_{g-1}} &= a_{f_{II} p}, \\
\ldots + \ldots &= \ldots \\
a_{f_{g-1} x_{g-1}} \cdot \beta_{x_{g-1}} &= a_{f_{g-1} p}.
\end{aligned} \tag{12}$$

The coefficients for the X's for use in the above equations are the diagonal factor loadings of the first g-1 X-variables as shown in Table III. The $a_{f_s p}$'s are the solution to the simultaneous equations solved in Table VI. The results of step 7 are shown in Table VII.

Table VII

The Standard Score Beta Coefficients (β_{X_i}'s) for the g-1 X-Variables Which Perfectly Predict the Vector Through the X Space Capable of Explaining the Most Variance

β_{X_1}	β_{X_2}	β_{X_3}	β_{X_4}	β_{X_5}	β_{X_6}
-.38191	.46194	.37010	-.48376	-.61559	.36691

Step 8.

Obtain the raw-score beta coefficients for the g-1 X-variables upon which the diagonal factors are based. This is accomplished by taking the mean of the Principal Axis Vector to be zero and taking its variance to be unity. Raw score beta for the $_i$th X-variable becomes

$$b_{X_i} = \beta_{X_i} \cdot \frac{\sigma_p}{\sigma_{X_i}} = \beta_{X_i} \cdot \frac{1.00}{\sqrt{\frac{n_{X_i}(N - n_{X_i})}{N}}},$$

$$= \beta_{X_i} \cdot \frac{N}{\sqrt{n_{X_i}(N - n_{X_i})}} \quad . \tag{13}$$

Obtain the constant to be added (A) by the equation

$$A = \bar{P} - \left(\sum^{g-1} b_{X_i} \cdot \bar{X}_i\right) = 0.0 - \sum^{g-1} b_{X_i} \cdot \frac{n_{X_i}}{N}$$

$$= \frac{-\sum^{g-1} b_{X_i} \cdot n_{X_i}}{N} \quad . \tag{14}$$

Table VIII

The Raw Score Beta Coefficients (b_{X_i}'s) for the g-1 X-Variables and the Constant to be Added (A)

b_{X_1}	b_{X_2}	b_{X_3}	b_{X_4}	b_{X_5}	b_{X_6}	A
-.90867	3.275	.82315	-1.5188	-.16125	2.30387	.06111

Step 9.

The values in Table VIII give weights to use in a prediction equation of the form

$$P = b_{x_1} \cdot X_1 + b_{x_2} \cdot X_2 + \ldots + b_{x_i} \cdot X_i + \ldots + b_{x_{g-1}} \cdot X_{g-1} + A. \qquad (15)$$

If we consider the predicted score of a person in the i^{th} X category, the only X-variable to have a nonzero score will be X_i, which will have a value of 1.0. Therefore, the predicted score (i.e., location of case on the Principal Axis vector) for a person who was in the i^{th} category of the X-variable will be

$$P_{x_i} = b_{x_1}(0) + b_{x_2}(0) + \ldots + b_{x_i}(1) + \ldots + b_{x_{g-1}}(0) + A,$$

which simplifies to

$$P_{x_i} = b_{x_i} + A. \qquad (16)$$

For a case which was in the g^{th} category of X, the predicted score will be simply $P_{x_g} = A$. These scores represent the locations of the various X categories on a vector through the X-space which must be optimally related to the variances of the Y-variables. The mean of the scores will be zero and the standard deviation will be one. For this reason we refer to these predicted scores (P_x's) as the z-weights (i.e., $z_{x_i} = P_{x_i}$). Table IX shows the computed z-weights based on Eq. (16).

Table IX

The Computed z-Weights for the g X-Variables

z_{x_1}	z_{x_2}	z_{x_3}	z_{x_4}	z_{x_5}	z_{x_6}	z_{x_7}
-.84755	3.78582	.88426	-1.45766	-.10014	2.36498	.06111

Step 10.

Inasmuch as the X categories have been ordered along a single continuum, we may consider X to be an interval-type variable. The optimal set of weights for the Y categories may be obtained by Wherry's (1) multiserial equation which states

$$b_{y_i} = \sum^{n_j} X/n_j,$$

which, for our purposes, may be rewritten as

$$b_{y_j} = \frac{\sum^{g} (z_{x_i} \cdot C_{ij})}{n_j} \quad , \tag{17}$$

where:

b_{y_j} = the multiserial weight for the j^{th} Y category,

z_{x_i} = the x-weight for the i^{th} X category,

c_{ij} = the number of cases in the i^{th} X category which were also in the j^{th} Y category, and

n_j = the total number of cases in the j^{th} Y category.

The multiserial weights for the Y cctegories, as computed by Eq. (17), are shown in Table X.

Table X

The Multiserial Weights for the Y Categories

b_{y_1}	b_{y_2}	b_{y_3}	b_{y_4}	b_{y_5}	b_{y_6}	b_{y_7}	b_{y_8}
-.86933	1.01338	-.39234	1.76070	-.65546	.03737	.72035	.26810

Step 11.

Obtain the K-coefficient by the equation

$$K_{xy} = \sqrt{\frac{\sum^{h} b_{y_j}^2 \cdot n_j}{N}} \tag{18}$$

For the present problem K_{xy} has a value of .56219.

Step 12.

This final step is optional but may be accomplished if z-weights are desired for the Y variables instead of multiserial weights. The z-weights, when applied to the entire sample, will yield a mean of zero and a standard deviation of unity. The z-weights for the h Y-variables are obtained by applying the equation

$$z_{y_j} = b_{y_j} / K_{xy} \quad . \tag{19}$$

The calculated z-weights for the Y-variables are shown in Table XI.

Table XI

The z-Weights for the h Y-Variables

z_{y_1}	z_{y_2}	z_{y_3}	z_{y_4}	z_{y_5}	z_{y_6}	z_{y_7}	z_{y_8}
-1.54632	1.80256	-.69788	3.13186	-1.165992	.06630	1.28133	.47688

Proof that K_{XY} is a Pearson-Moment Correlation Coefficient

Using the z-weights for the g X-categories and the h Y-categories will yield means of zero and standard deviations of unity for both X and Y. Under such circumstances the Pearson product-moment correlation coefficient may be computed as

$$r_{XY} = \frac{\sum^{N} z_x z_y}{N} \quad .$$

The above equation may be rewritten as

$$r_{XY} = \frac{\sum^{h}\sum^{g}(z_{x_i} \cdot z_{y_j} \cdot c_{ij})}{N} = \frac{\sum^{h}\left(z_{y_j} \sum^{g}(z_{x_i} \cdot c_{ij})\right)}{N} \qquad (20)$$

However, from Eq. (17) we know

$$b_{y_j} \cdot n_j = \sum^{g}(z_{x_i} \cdot c_{ij}) \quad ,$$

therefore, substituting this value into Eq. (20) we obtain

$$r_{XY} = \frac{\sum^{h} z_{y_j} \cdot b_{y_j} \cdot n_j}{N} \quad . \qquad (21)$$

Now, substituting Eq. (19) into Eq. (20) we obtain

$$r_{XY} = \frac{\sum^{h} b_{y_j}^2 n_j}{N \cdot K_{XY}} \quad . \qquad (22)$$

From Eq. (18) we see that

$$K^2_{XY} = \frac{\sum^{h} b^2_{y_j} \cdot n_j}{N} ,$$

therefore, substituting this value into Eq. (22) we see

$$r_{XY} = \frac{K^2_{XY}}{K_{XY}} = K_{XY} ,$$

which proves the K-coefficient is, in fact, a Pearson-type measure of correlation for categorical data.

The data used in the above example were obtained by randomly combining various adjacent columns and various adjacent rows of some "interval" type data on "height of fathers" (X) and "height of sons" (Y) as found on page 117 of McNemar (9). The data were combined as shown in Figure 1.

The z-weights derived by the procedure described in this paper, when applied in subsequent samples, would not be expected to yield a correlation this high. A paper on the expected "shrinkage" of the K-coefficient is being prepared as well as a paper on testing the significance of the K-coefficient. Until the latter paper is published users of the K-coefficient may use the χ^2 -test associated with the contingency coefficient to decide if obtained K-coefficients differ significantly from zero.

Initial investigations of the distribution of K-coefficients indicate that a K-coefficient may be less than, as well as greater than, the corresponding C-coefficient.

The obtaining of a K-coefficient, especially when the number of rows and columns is large, is obviously a tedious process and one which should be handled by computer rather than desk calculator. The technique has been programmed for the IBM 1620 computer and will handle problems where $g + h \leq 20$. In spite of the complexity of the technique, the K-coefficient seems far more acceptable as the proper measure of relationship between categorical variables than the older contingency coefficient.

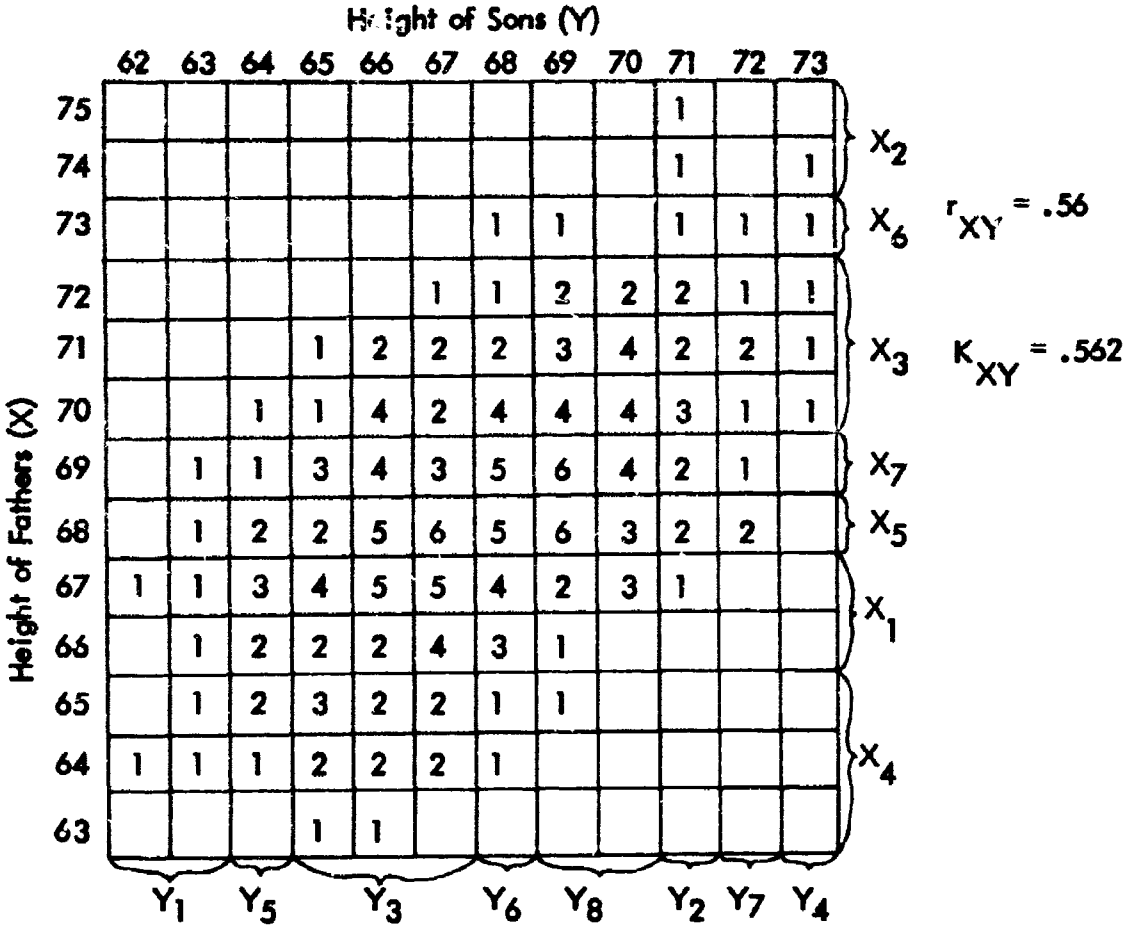

Figure 1. The Original Scatterplot of the Data from which Table I was Obtained

REFERENCES

1. Wherry, R.J., Sr., Maximal weighting of qualitative data. Psychometrika, 9: 263-266, 1944.

2. Bottenberg, R.A., and Ward, J.H., Jr., Applied multiple linear regression. Technical Documentary Report PRL-TDR-63-6. Lackland AFB, Texas: Personnel Research Laboratory, Aerospace Medical Division, AF Systems Command, 1963.

3. Wherry, R.J., Jr., Toward an optimal method of equating subgroups composed of different subjects. Monograph Series, No. 9. Pensacola, Fla.: Naval School of Aviation Medicine, 1964.

4. Hotelling, H., The most predictable criterion. J. educ. Psych., 16:139-142, 1935.

5. Fisher, R.A., Statistical Methods for Research Workers. 7th ed., Edinburgh: Oliver and Boyd, 1938.

6. McKeon, J.J., Canonical analysis: Some relations between canonical correlation, factor analysis, discriminant function analysis, and scaling theory. Public Health Service Training Grant 2M-6961 and Nonr 1834(39). Urbana, Ill.: University of Illinois, Department of Psychology, July, 1962.

7. Fruchter, B., Introduction to Factor Analysis. Princeton, N.J.: D. Van Nostrand Co., Inc., 1954.

8. Hotelling, H., Analysis of a complex of statistical variables into principle components. J. educ. Psych., 24:417-441 and 498-520, 1933.

9. McNemar, Q., Psychological Statistics. 3rd ed. New York: John Wiley and Sons, Inc., 1962.

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) U.S. Naval School of Aviation Medicine Pensacola, Florida	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

THE K-COEFFICIENT, A PEARSON-TYPE SUBSTITUTE FOR THE CONTINGENCY COEFFICIENT.

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (Last name, first name, initial)

Wherry, Robert J., Jr., and Lane, Norman E.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
25 May 1965	15	9

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO. MR005.13-3003	NSAM - 929
c. Subtask 1	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) 43
d.	

10. AVAILABILITY/LIMITATION NOTICES

Qualified requesters may obtain copies of this report from DDC. Available, for sale to the public, from the Clearinghouse for Federal Scientific and Technical Information, Springfield, Virginia, 22151.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

The Pearson product-moment correlation coefficient (r) is recognized as the basic equation for relationship. Well-known and widely used derivatives of the Pearson equation include the point-biserial correlation ($r_{pt.\ bis.}$), the Spearman rank-difference correlation (ρ), the fourfold point correlation ($_{\varphi}$), and the correlation ratio (η).

Expression of relationships between categorical variables has previously been possible only by means of the contingency coefficient. This paper describes an extension of Pearson's basic equation to provide a measure of relationship between two categorical variables, the "K" coefficient, which avoids many of the disadvantages inherent in the contingency coefficient.

DD FORM 1 JAN 64 1473

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Correlation Correlation analysis Canonical correlation Factor analysis Statistics Statistical analysis Qualitative data						

INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.